Lending Club Case Study

SHRUTI CHOUDHARY

JEYASHREE M

Problem Statement

When the lending club company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

The company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Business Objective

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

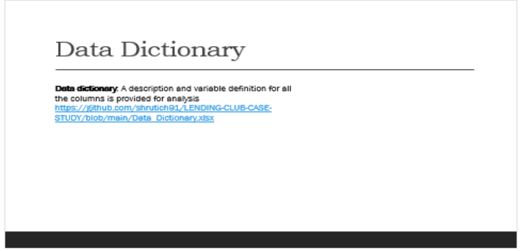
If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicant's using EDA is the aim of this case study.

Data Understanding

[Loan dataset]

Rows	Columns
39717	111

ADDITIONAL DETAILS ON DATA DICTIONARY



Column Name	Description
Loan_Amount	Amount requested by applicant
Funded Amount	The total amount commited to that loan at that point of time
Term	Loan duration term in months
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
grade	Lending club assigned grade
sub_grade	Lending club assigned sub-grade (within grade)
emp_title	The job title supplied by the Borrower when applying for the loan.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
home_ownership	The home ownership status provided by the borrower during registration.
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income was verified by Lending club
issue_d	The month which the loan was funded
loan_status	Current status of the loan. It can be fully paid or charged off or current
purpose	A category provided by the borrower for the loan request.
title	The loan title provided by the borrower
addr_state	The state provided by the borrower in the loan application
dti	A ratio calculated using the borrowers total monthly debt payments on the total debt obligations
revol_util	Revolving line utilization rate
pub_rec_bankruptcies	Number of public record bankruptcies
recoveries	post charge off gross recovery

Steps involved in Analysis

1 Importing

Importing Libraries and Reading Data

- Import all required libraries
- Example: pandas, numpy, Math, matplotlib, plotly, seaborn, wordcloud

2 Data Cleaning –null values

Null values

- removing columns with all null values

-removing columns with all same values

- removing columns with null values > 60%

- removing columns with distinct values

- removing columns with customer behaviour

- Treating null values if required with mode/mean/median -

3 Data Cleaning – outliers, transformation

Data Types Analysis, Data Formatting, Transformation

- Convert the datatype of columns for better analysis
- Create derived columns bot better analysis
- Rename columns to meaningful names

Treating outliers

- Study skewness, kurtosis

- Identify outliers and treat outliers if required

4 Data Summary Data Integrity Checks

Data Types Analysis, Data Formatting, Transformation

- Convert the datatype of columns for better analysis
- Create derived columns bot better analysis

Treating outliers

- Study skewness, kurtosis

- Identify outliers and treat outliers if required

5 Univariate, Bivariate, Multivariate Analysis

Univariate Analysis

- Analyze the distribution of all necessary columns

Bivariate Analysis

- Study the target variable against all variables
- Identify trends and
- observations

Multivariate Analysis

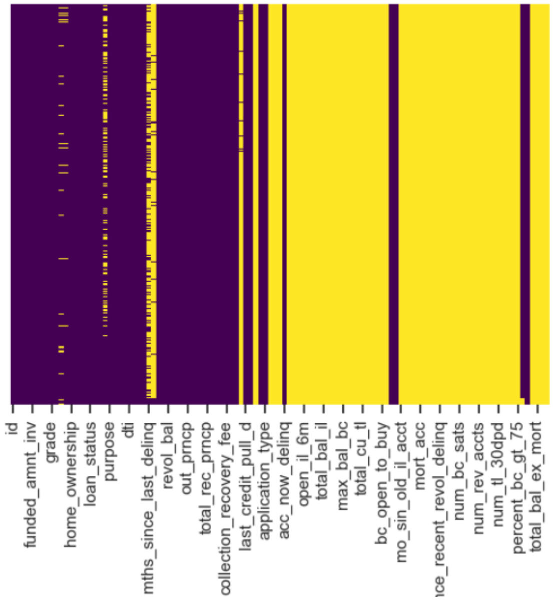
- Identify the drivers and create multivariate analysis to reach conclusions

Importing the dataset

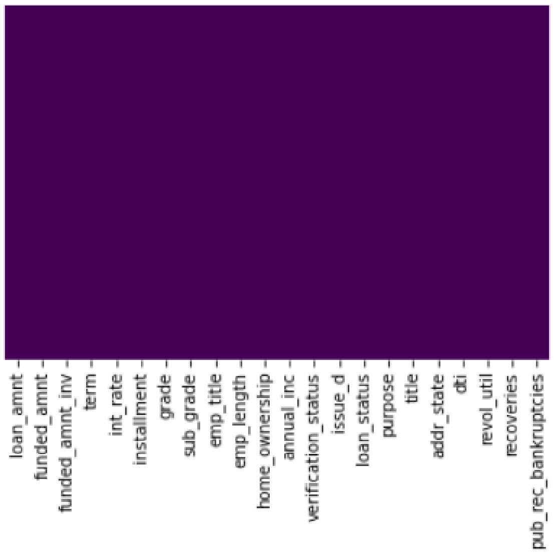
- The dataset is imported into pandas dataframe
- Total rows: 39717
- Total columns – 111
- Python Libraries used for analysis:
  - Pandas for data analysis and transformation
  - Seaborn, matplotlib, wordcloud, plotly for visualizations

Date Manipulation Before & After

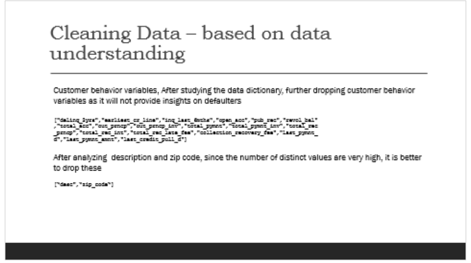
BEFORE MANIPULATION
OF NULL VALUES



AFTER MANIPULATION
OF NULL VALUES

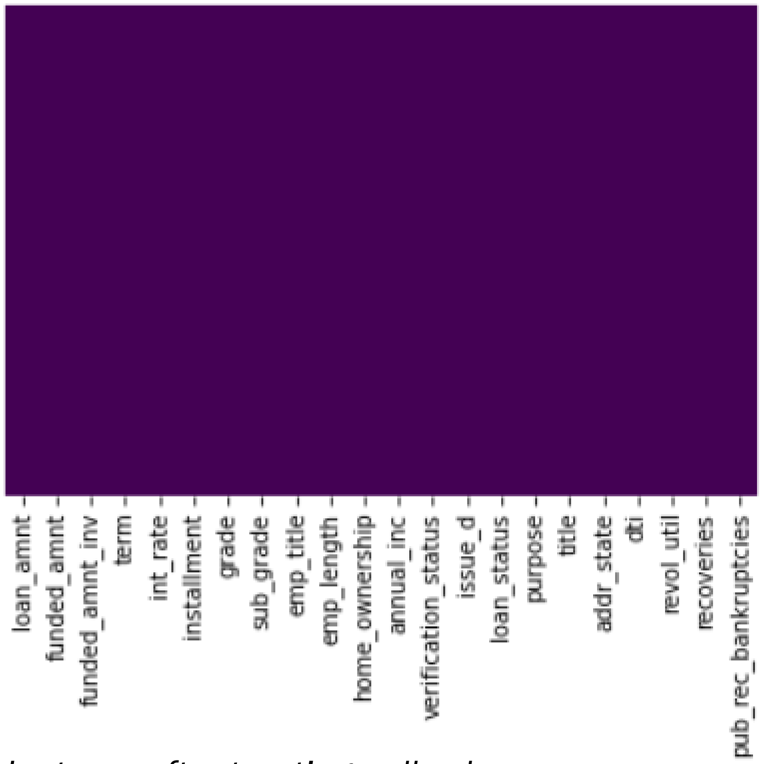


DETAILED DATA
CLEANING EXPLAINED



Treating null values

Column	Null value treatment
pub_rec_bankruptcies	Mode
emp_title	None
emp_length	0 (As numerical analysis is easier for this)



Heatmap after treating null values

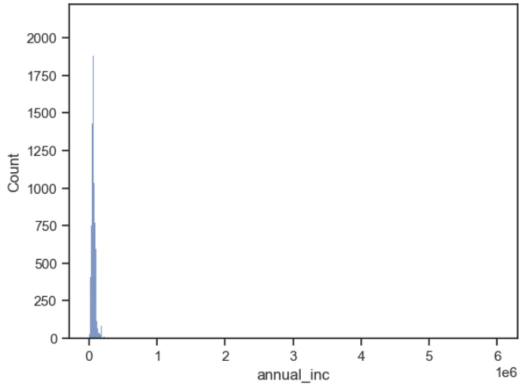
Data Types Analysis, Data Formatting, Transformation

Feature name	Modified Feature name	Type of conversion	Explanation
int_rate	Int_rate	Object to float	Removal of % from data will make it float, easier for analysis
emp_length	emp_length_year	Object to int	Easier for bucketing
issue_d	Issue_d_month, issue_d_year	Extracted month and year	Easier for analysis on month, year
revol_util	revol_util_rate	Object to float	Removal of % from data will make it float, easier for analysis

Other features that are transformed to derived columns using bucketing logic during bivariate analysis- annual_inc,dti, loan_amnt , emp_title, installment , int_rate_bucket

Treating outliers

After analysis of all numerical columns, the following columns had outliers



Skewed annual income before treatment

Feature	Threshold (upper threshold)	% of data reduced	Explanation
loan_amnt	29250	3.1%	Some higher loan amounts, can create bias in the analysis.
annual_inc	139537.5	4.35%	People with very high income will create skewness in analysis.
installment	771	2.86%	Some installments may be very high, as people may have missed previous installments.

Data Integrity Checks and Summary

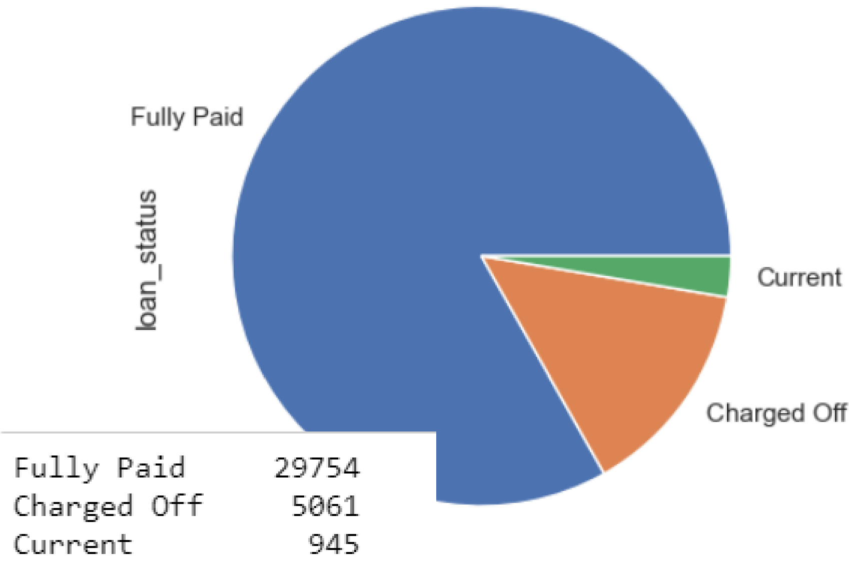
As per business understanding, loan_amnt >= funded_amnt >= funded_amnt_inv. Performed test to check data integrity. All records are valid.

Summary: Total 35760 rows, 26 columns are remaining for analysis

```
Data columns (total 26 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   loan_amnt             35760 non-null  int64
 1   funded_amnt           35760 non-null  int64
 2   funded_amnt_inv       35760 non-null  float64
 3   term                  35760 non-null  object
 4   int_rate              35760 non-null  float64
 5   installment           35760 non-null  float64
 6   grade                 35760 non-null  object
 7   sub_grade             35760 non-null  object
 8   emp_title             35760 non-null  object
 9   emp_length            34734 non-null  object
 10  home_ownership        35760 non-null  object
 11  annual_inc            35760 non-null  float64
 12  verification_status   35760 non-null  object
 13  issue_d               35760 non-null  object
 14  loan_status           35760 non-null  object
 15  purpose               35760 non-null  object
 16  title                 35749 non-null  object
 17  addr_state            35760 non-null  object
 18  dti                   35760 non-null  float64
 19  revol_util            35712 non-null  object
 20  recoveries            35760 non-null  float64
 21  pub_rec_bankruptcies  34734 non-null  object
 22  issue_d_year          35760 non-null  object
 23  emp_length_year       35760 non-null  int64
 24  revol_util_rate       35712 non-null  float64
 25  issue_d_month         35760 non-null  object
```

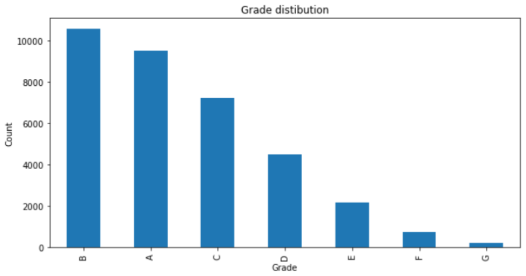
Univariate Analysis

Loan status



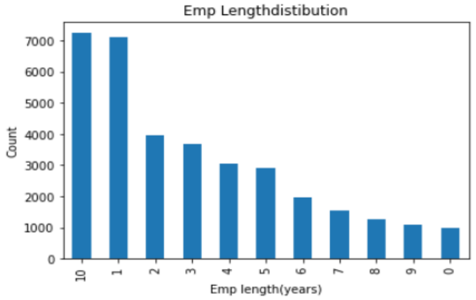
Dropping rows with loan status- current, as it is not useful for our analysis

Grade



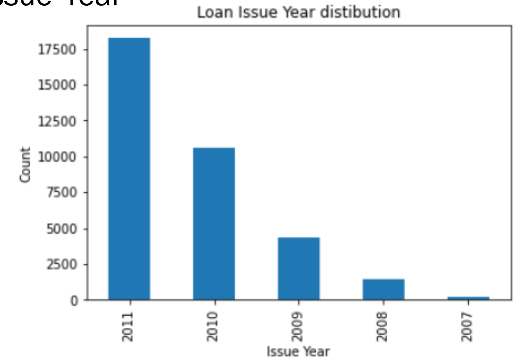
Most loan applicants have Grade, A, B

Employment length



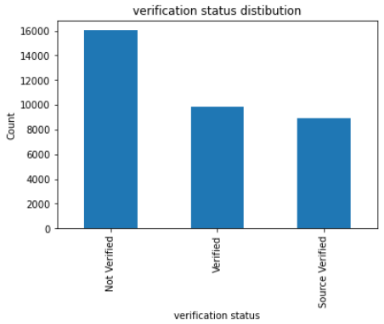
Most loan applicants are either 10+, Freshers.

Issue Year



Most loan applicants applied in 2011

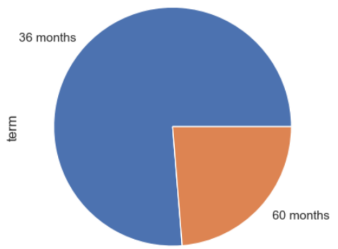
Verification_status



Most loan applicants have verification status as not Verified

Univariate Analysis

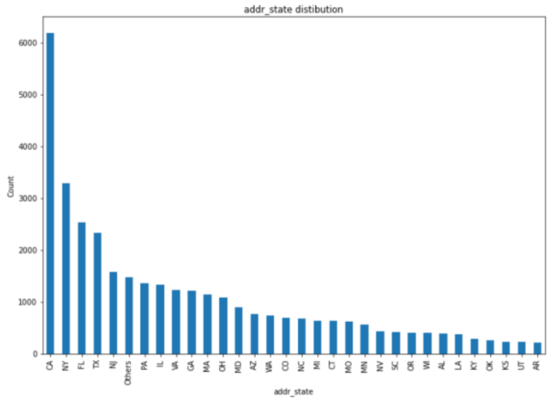
Term




```
36 months     26561
60 months      8254
```

Address State

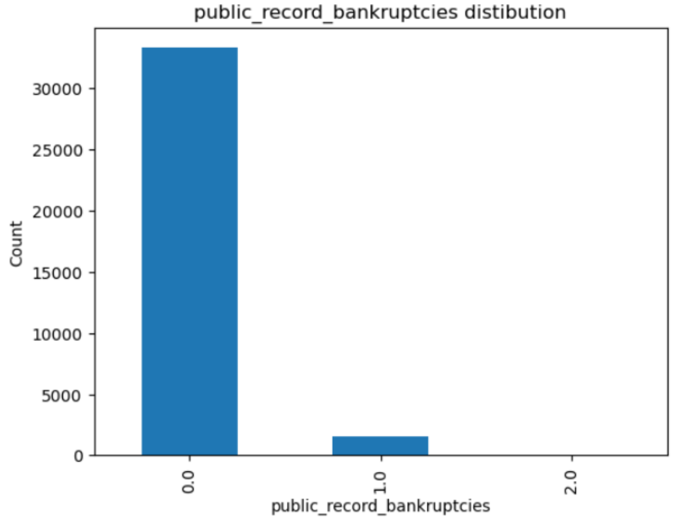
Most loan applicants are from CA



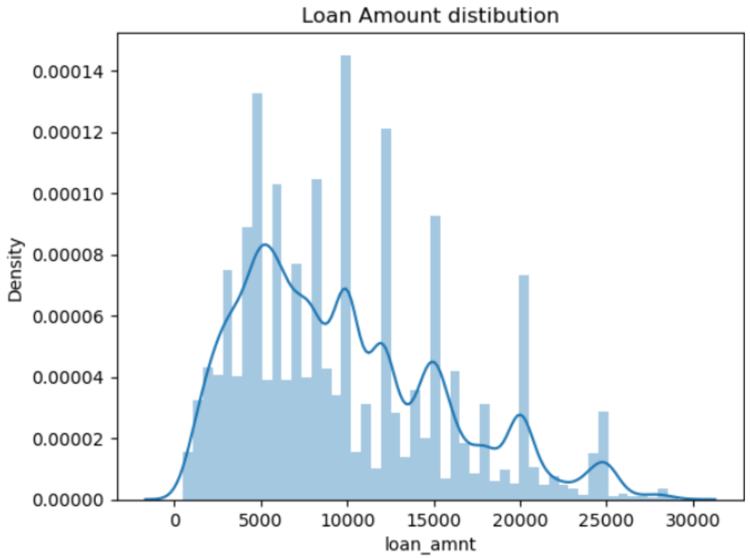
Word cloud on title shows debt consolidation as the highlight, since it is captured in purpose- dropping the column

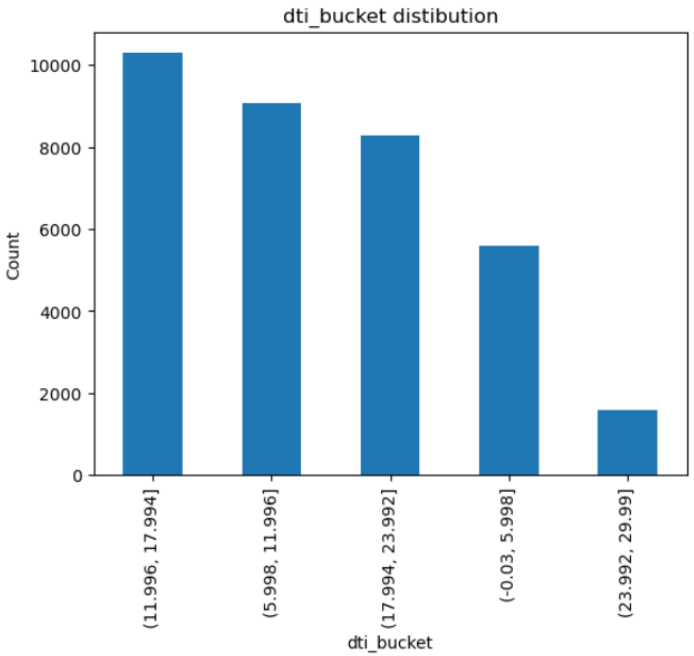
bankruptcies

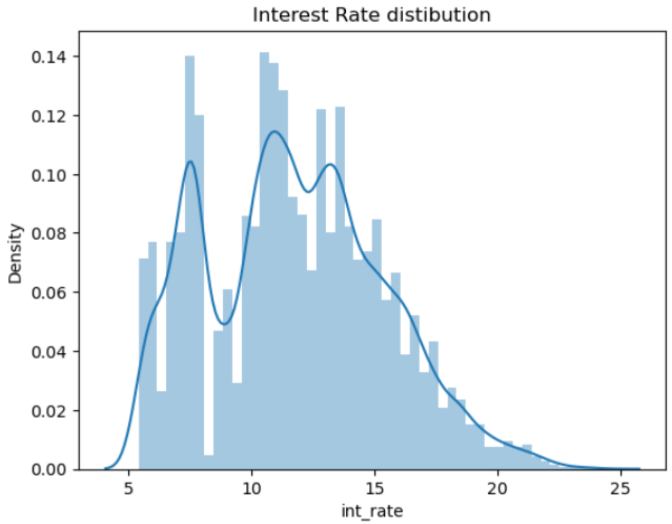
Most loan applicants have zero bankruptcy



Univariate analysis





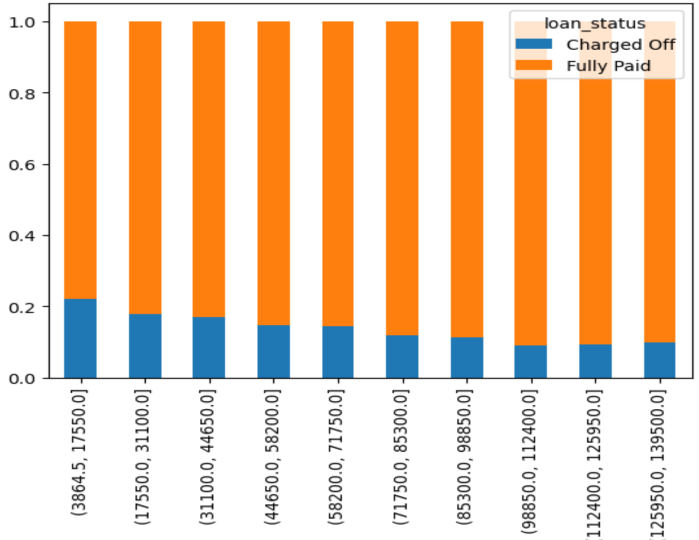


People are rounding off their loan amount to multiples of 1000s when applying for loan. Loan Amount has peaks at 5000, 10000, 15000, 20000, 25000s

Most applicants have dti bucket in the mod range

Interest rate has 3 peaks

Bi-Variate Analysis

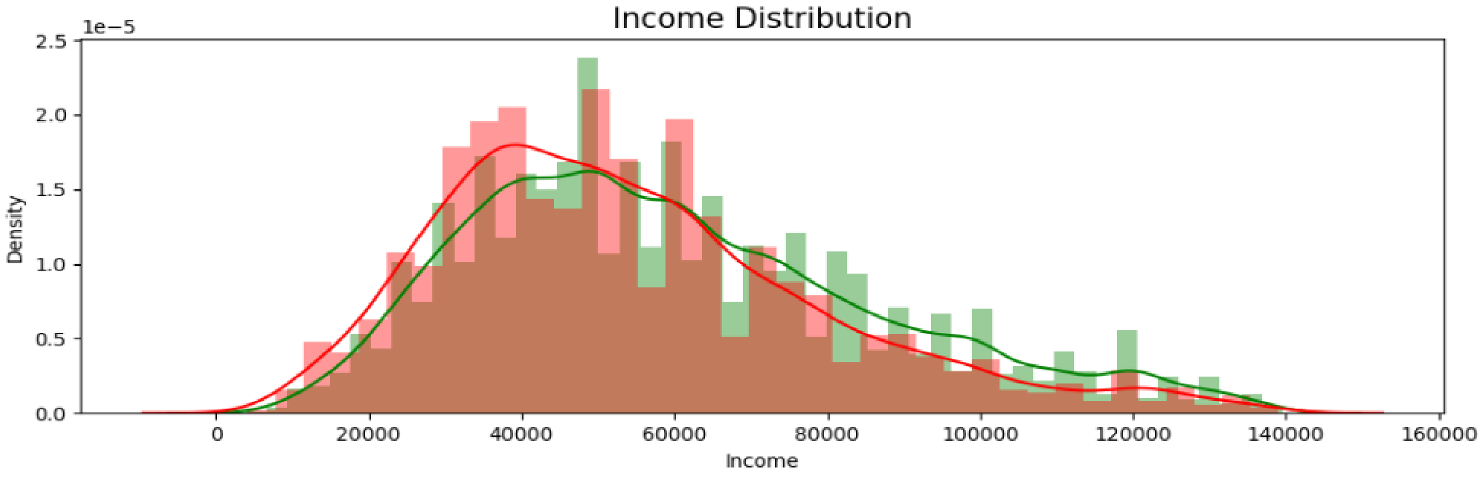


Loan status(ratio) Vs Income Bucket:

Trends- As income increases, %of charged off decreases

Observation - Income <60k has a higher chance of charged off

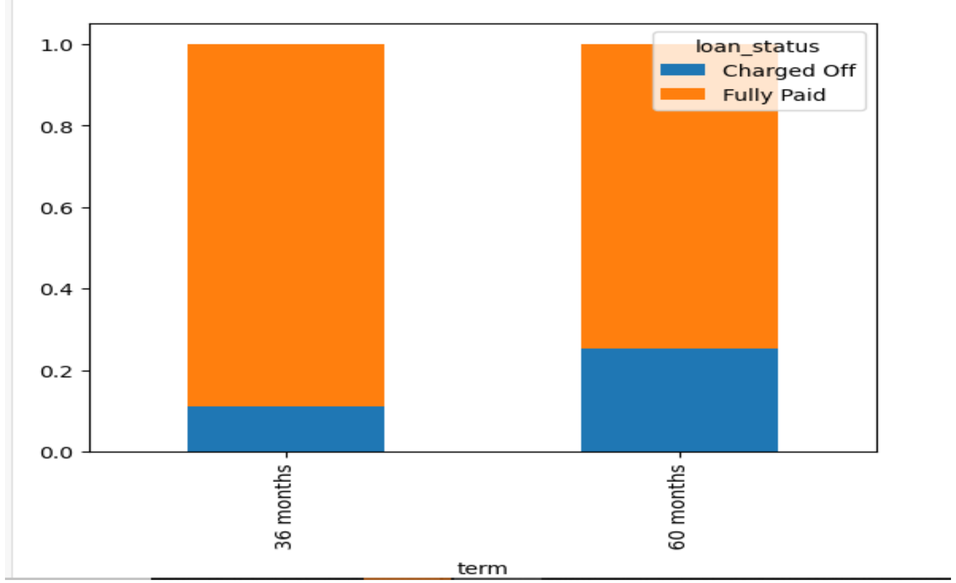
Observation : Till 60k income , charged off is outlying the fully paid, after 60k fully paid is more occurring



Bi-Variate Analysis

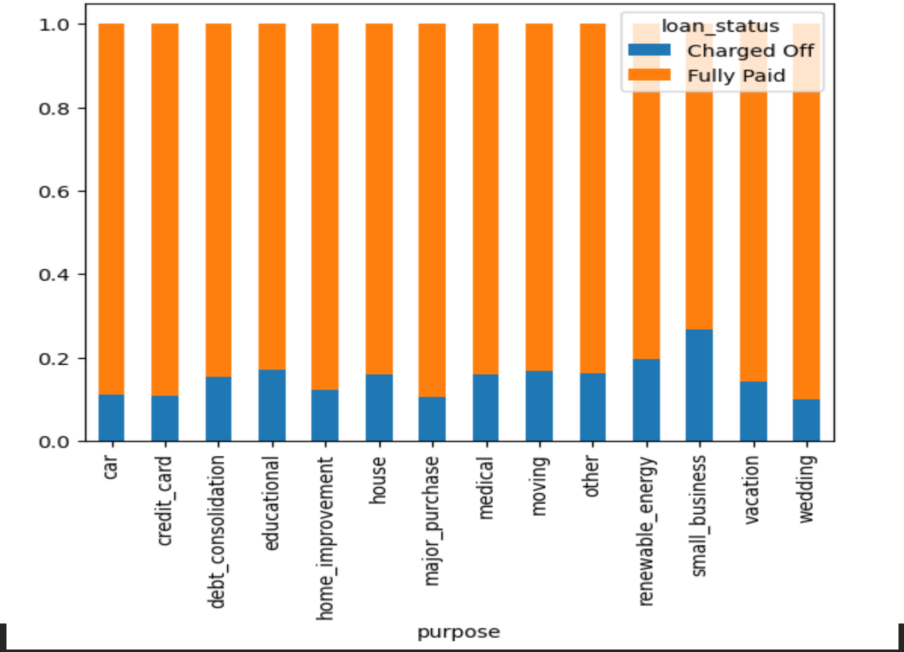
Loan status Vs Term

Insights : chance of charged off is double when term is 60



Loan status Vs Purpose

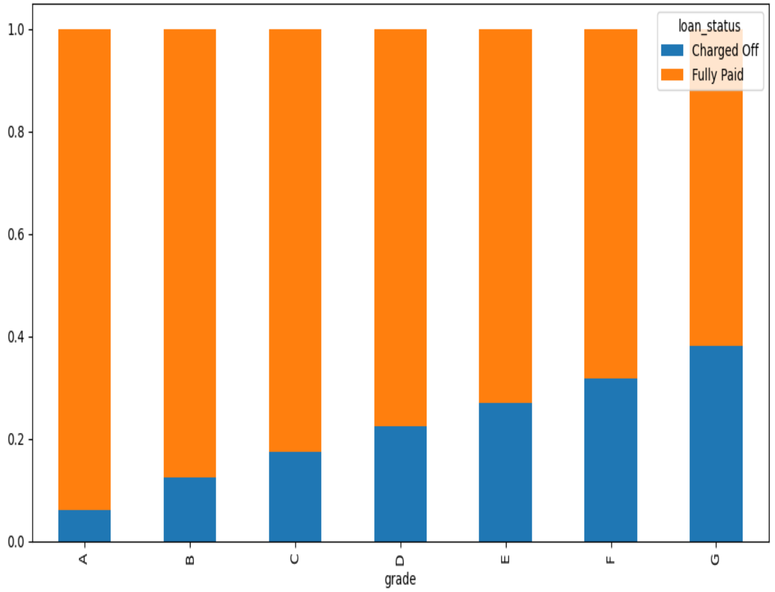
Insights: If purpose=Small business, it has highest(27%) chance of charged of



Bi-Variate Analysis

Loan status(ratio) Vs Grade

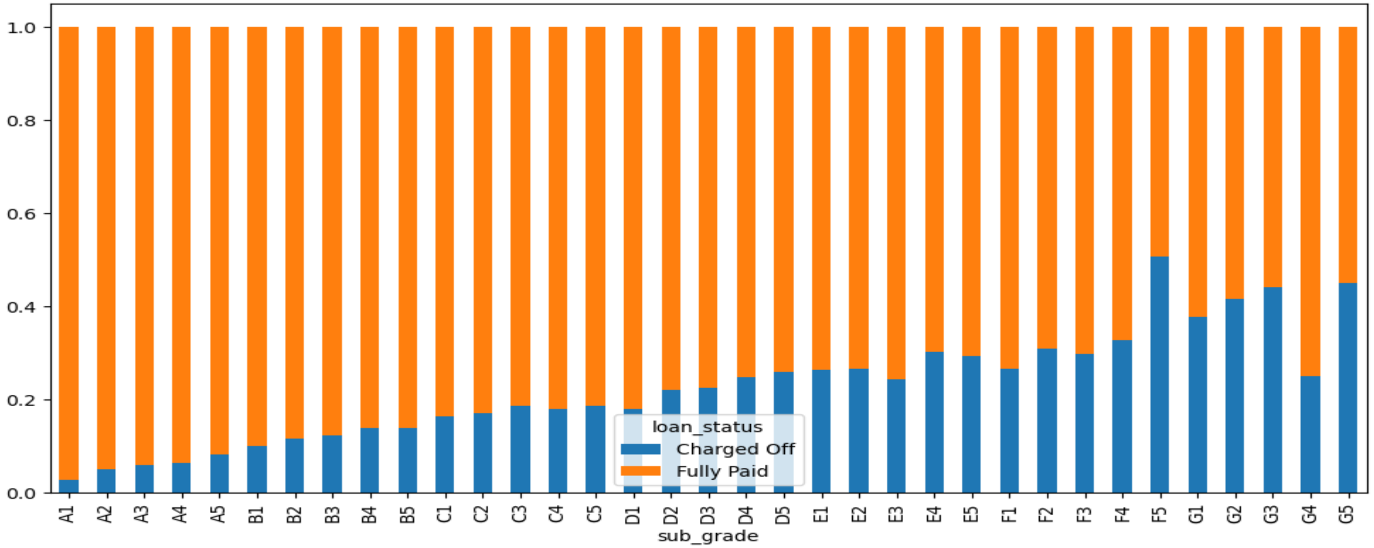
Insights: As grade increases(A to B to C), % charged off increases



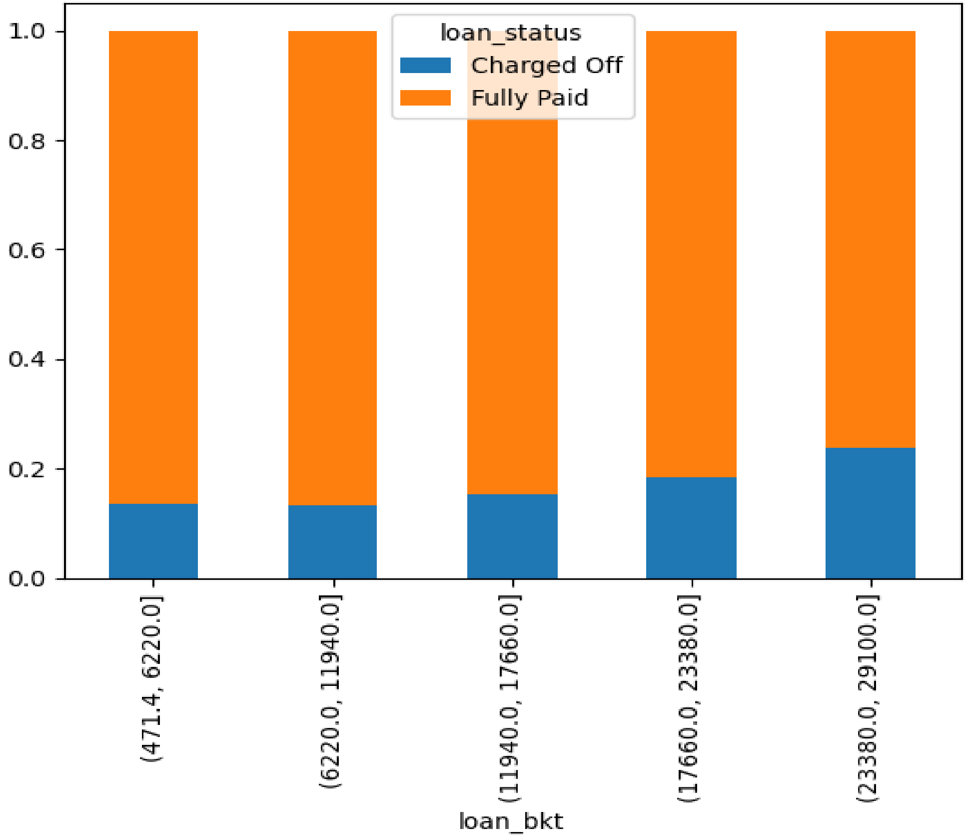
Loan status(ratio) Vs Subgrade

Insights: As subgrade increases, % charged off increases.

Also For F5 grade more charged off(50%) is observed



Bi-Variate Analysis

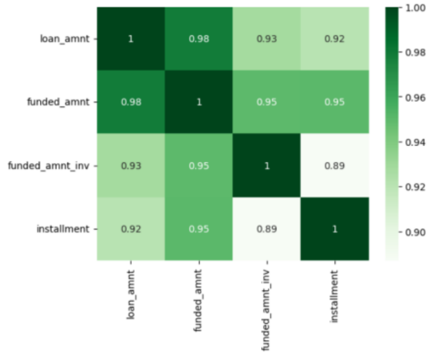


Loan Status(ratio) Vs Loan Amount:

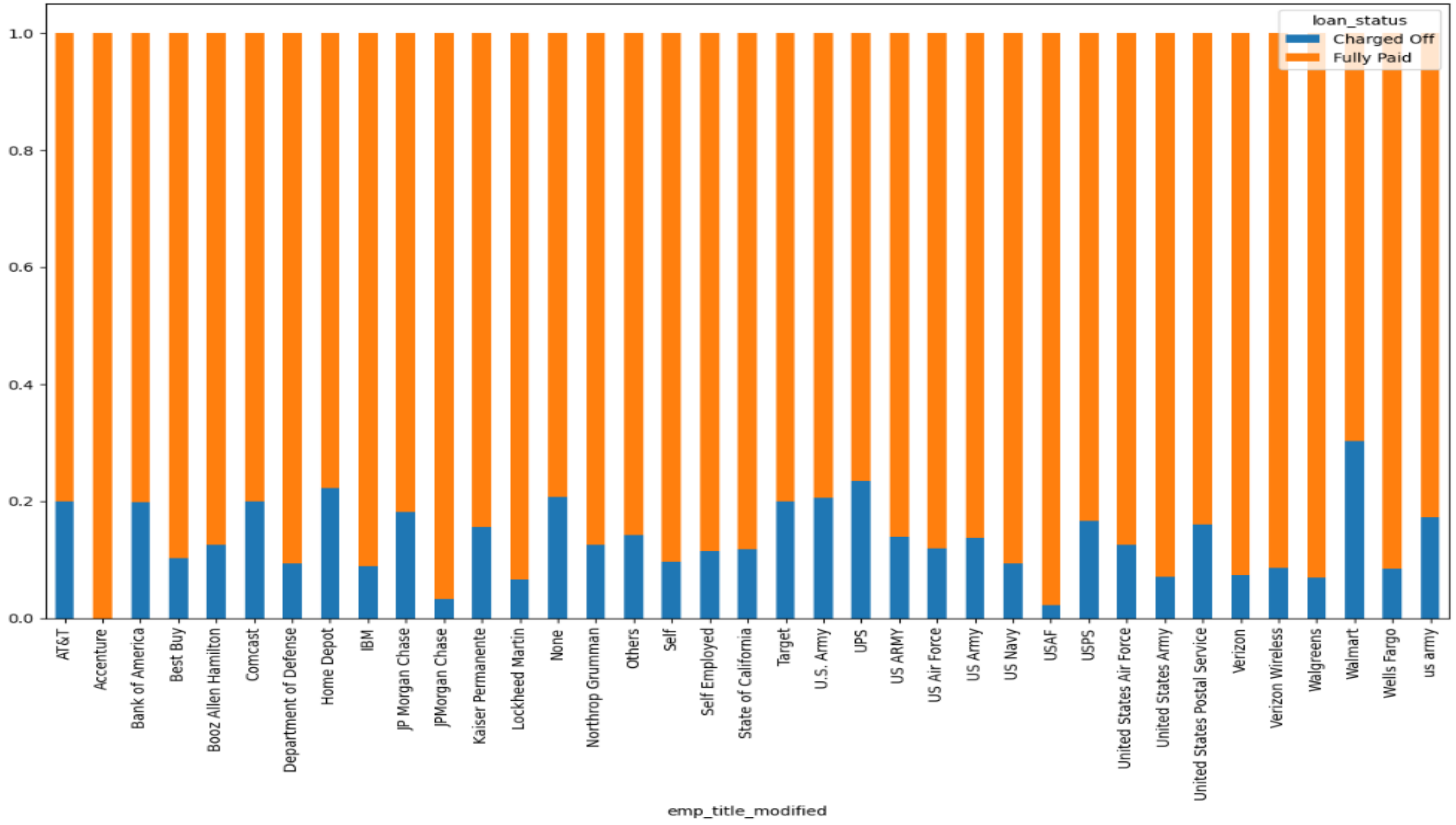
Insight: Loan bucket 23K-30k have highest percentage of chargeoff

Trend- As loan amount increases, % charged off increases

Because of positive **correlation**, loan amount , funded amount, funded_amnt_inv, installment will follow similar trend



Bi-Variate Analysis



Loan status Vs Title

Observation: title **Accenture has (0%) no charged off,** while Wallmart has highest Charged off

Observation: if employee title is null, higher chance of charged off

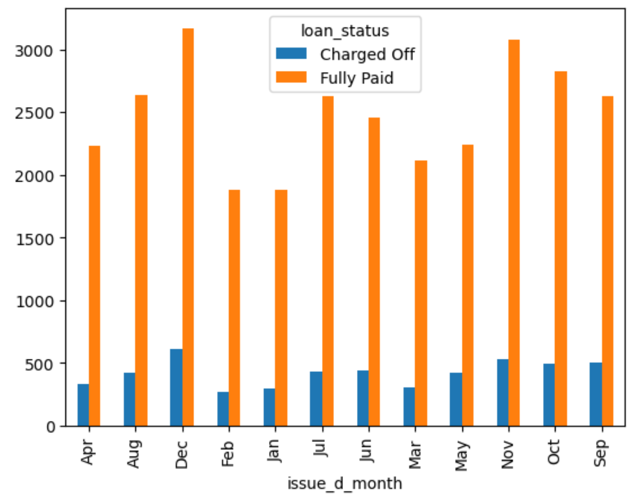
Since the data points for this conclusion is less, this cannot be a driver

Lending Club can tie up with Accenture for offers as there is a good success history

Bi-Variate Analysis

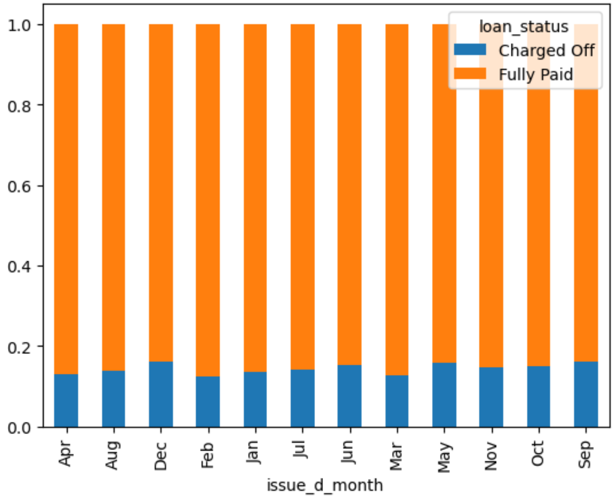
Loan Status(count) Vs Issue Month:

Trends- Most loan applicants were provided loan in Dec

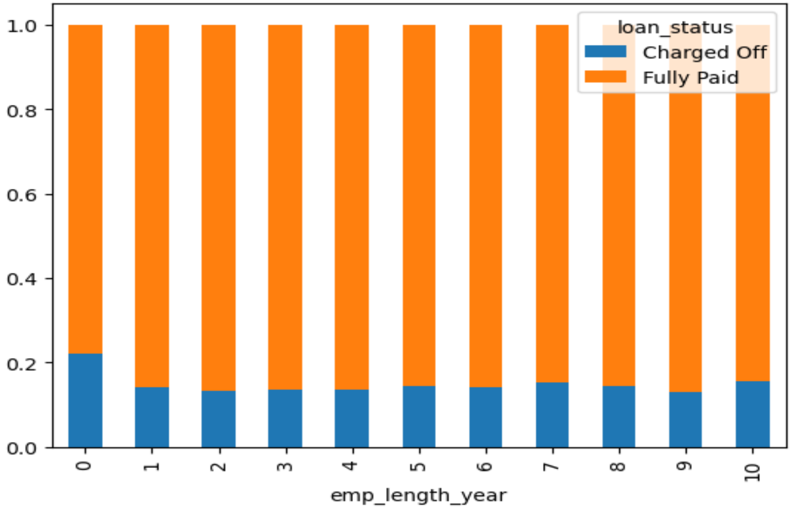


Loan Status(ratio) Vs Issue Month:

Trends- Month is not conclusive for loan status



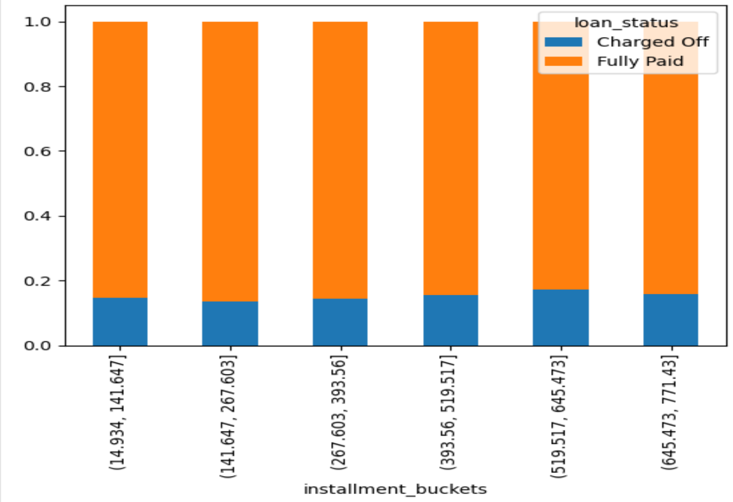
Bi-Variate Analysis



Loan Status(ratio) Vs Employment Length:

Observation - Employment length is not conclusive for any trend

Observation - If employment length in NA(plotted as 0) in graph, has higher(22%) chance of charged off as compared to others

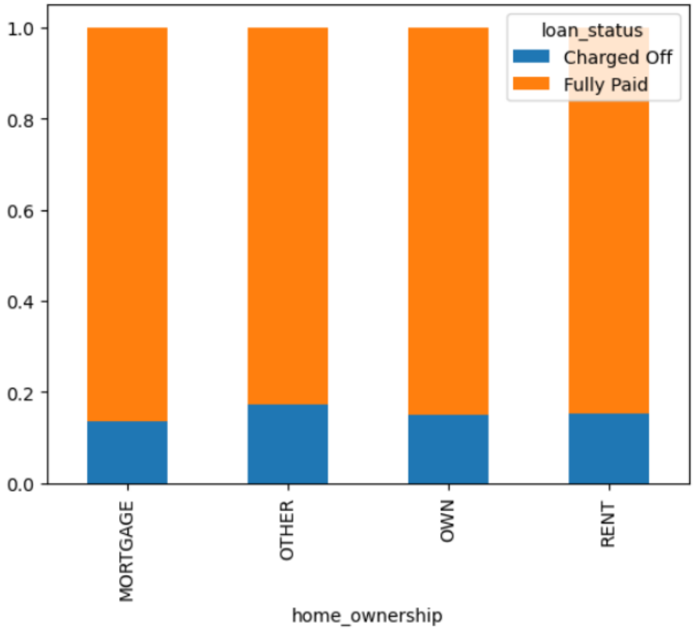


Loan status Vs Installment

Observation : It is not conclusive

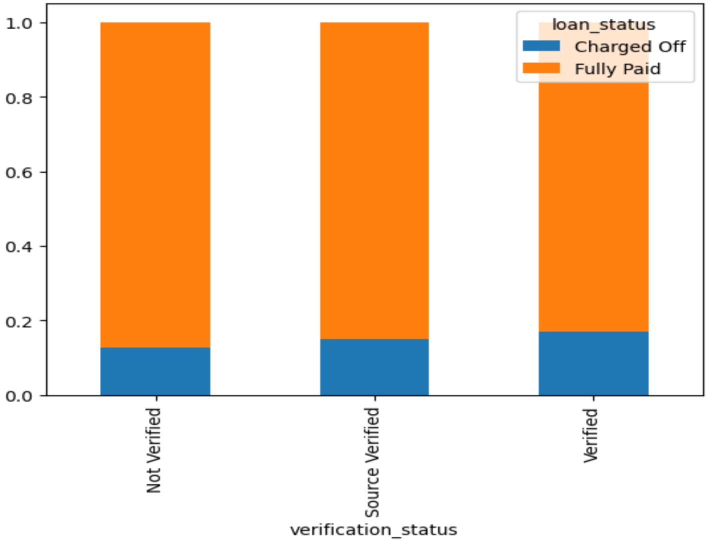
As installment increases there is no obvious evidence of charged off increase

Bi-Variate Analysis



Loan status(ratio) Vs Home ownership

Observation: This is not conclusive as there is no trend seen



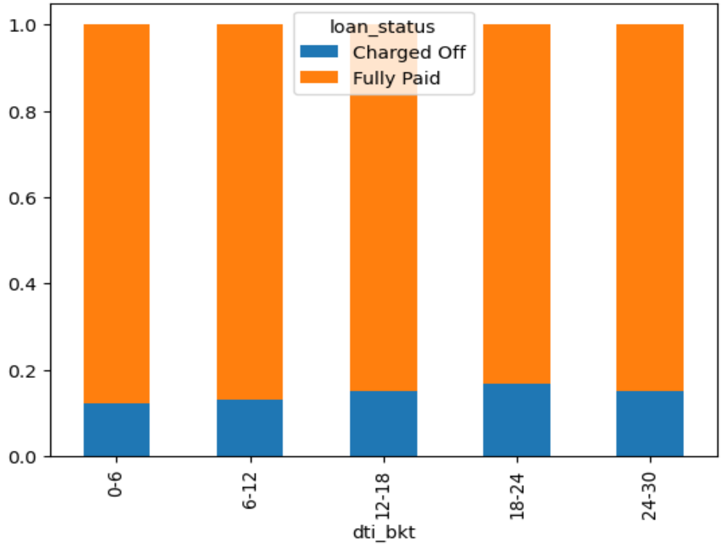
Loan status Vs Verification status

Observation: This is not conclusive as there is no trend seen

Bi-Variate Analysis

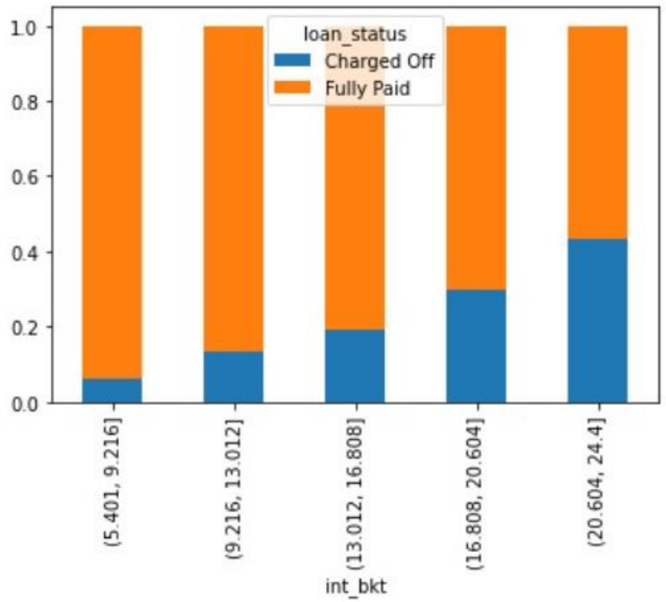
Loan status Vs Dti

Insights : dti is not showing a trend, 18-24% has comparatively higher chance of charged off

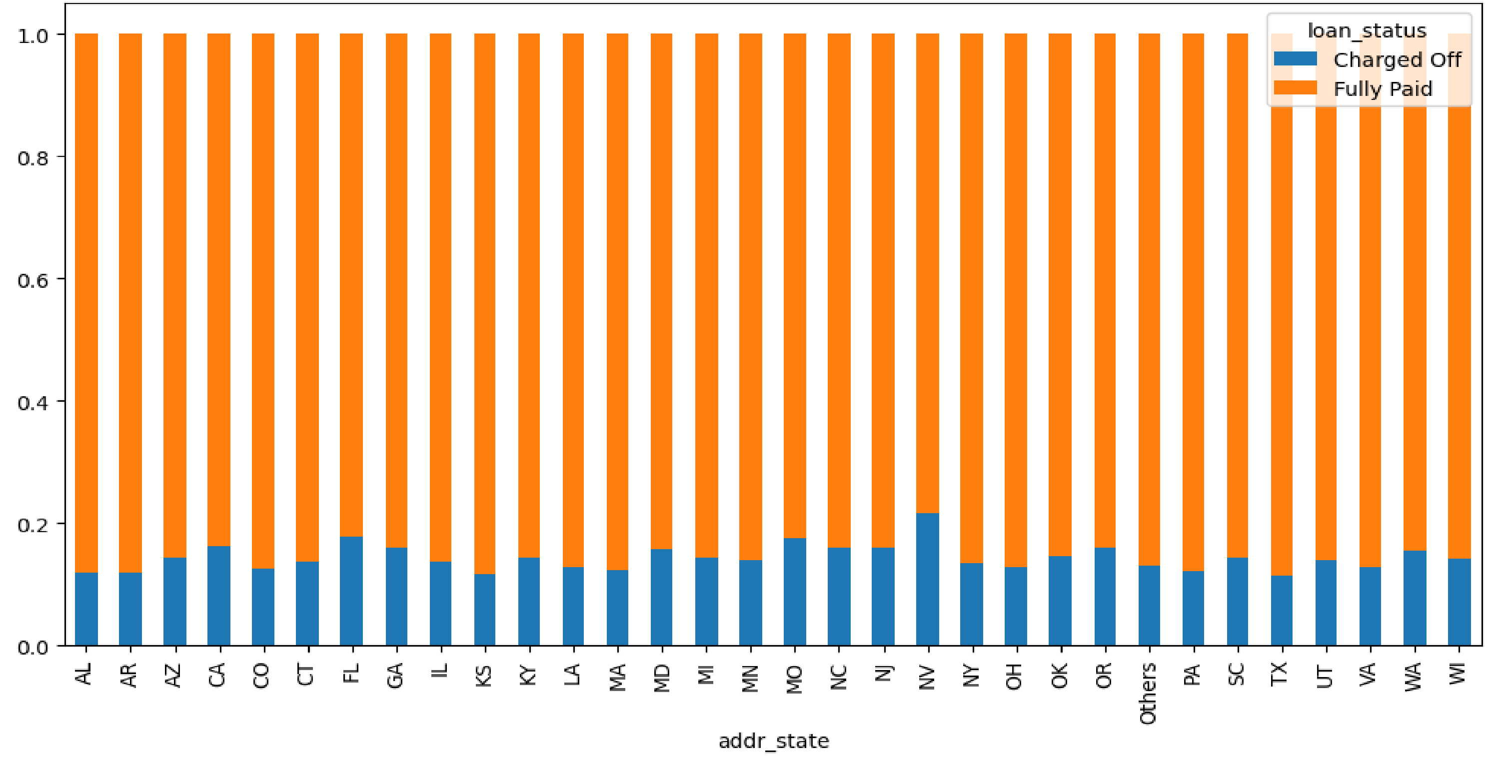


Loan status Vs Interest Rate

Insights : As interest increases, charged off increases



Bi-Variate Analysis



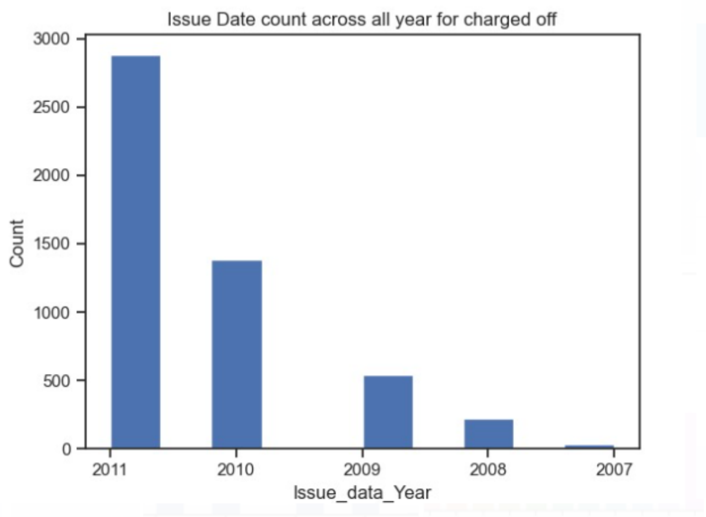
Loan status Vs addr state

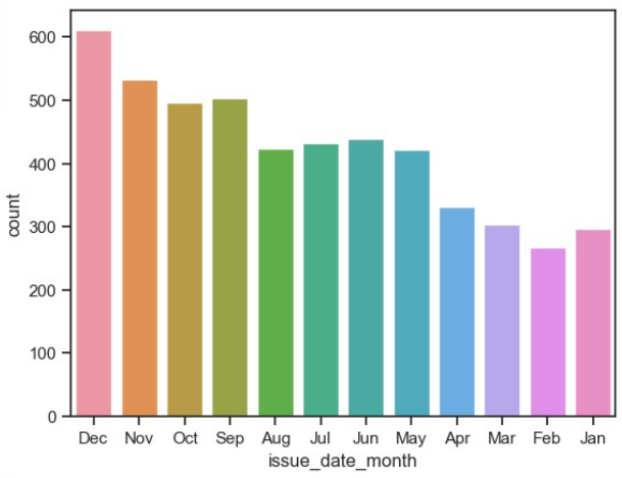
Insights: For State : NV proportion of charged off is more when compared to rest all

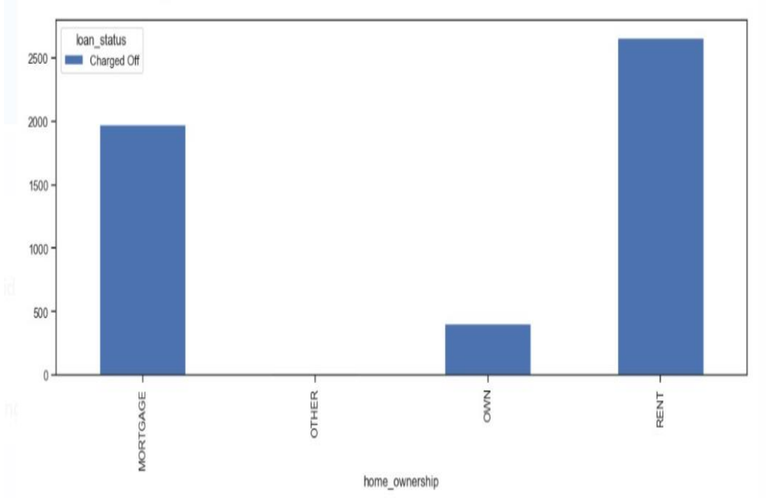
Note: All locations with very few data points(<200) are moved to Others category. If this is not done, state **NE has most charged off ratio**

Bivariate Analysis - Charged Off

For observations on charged off dataset, we have segregated charged off from the original dataset. Below patterns are not drivers but key observations on Charged off dataset – against issued year, month, home_ownership



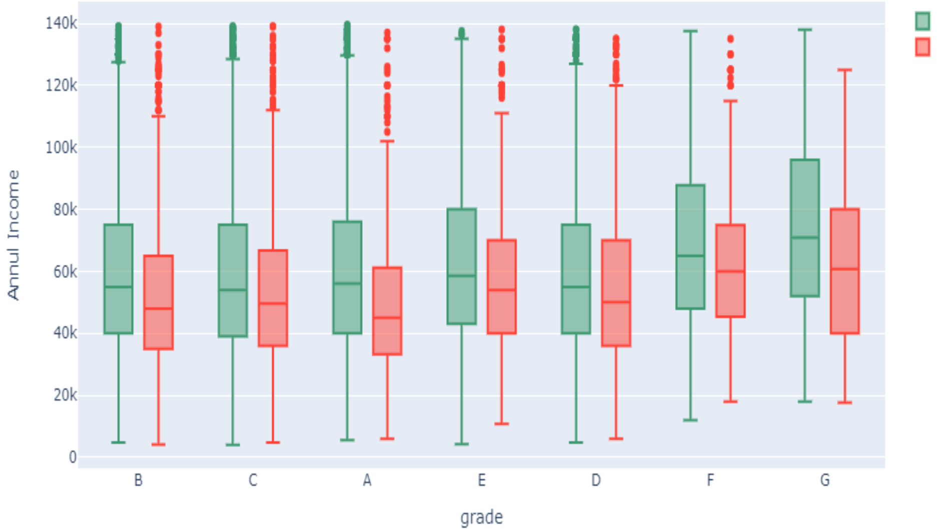


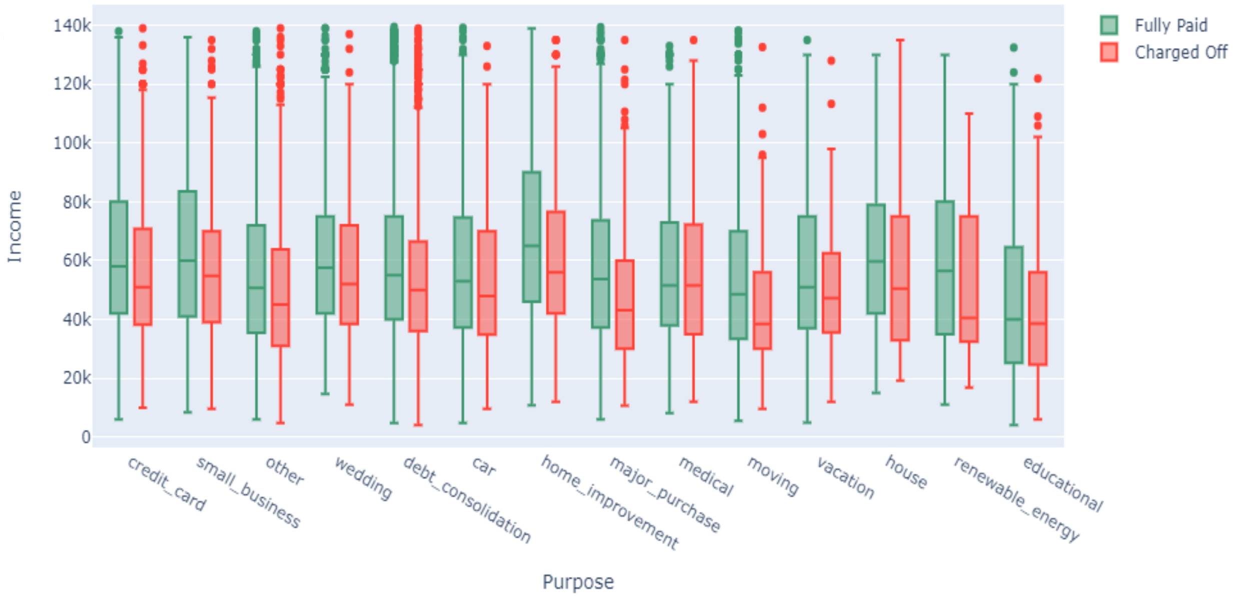


Multi Variate Analysis

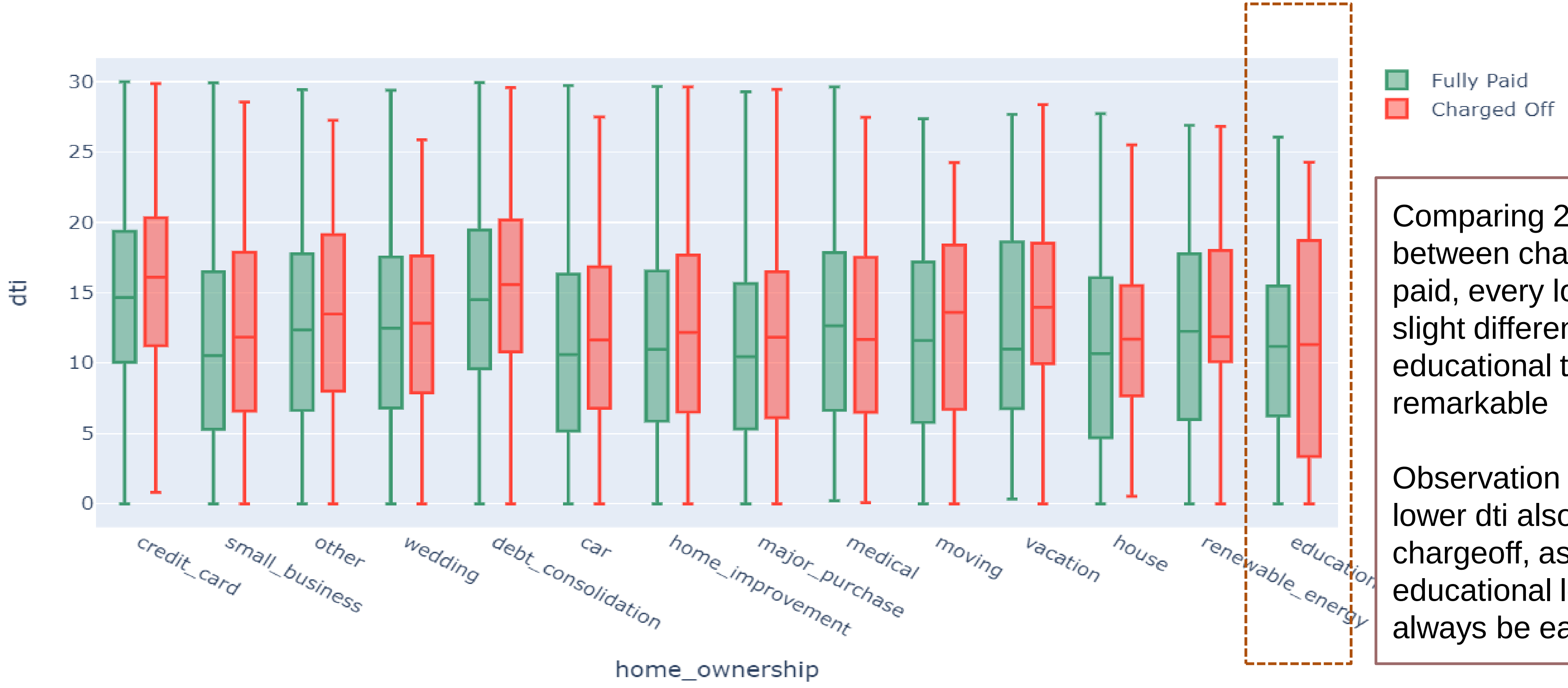
Observation: If loan applicant falls in Grade G and lesser annual Income chance of default is high

Observation: When purpose is home improvement/major purpose/moving/small business and income is lower respectively then higher chance of charged of





Multi Variate Analysis

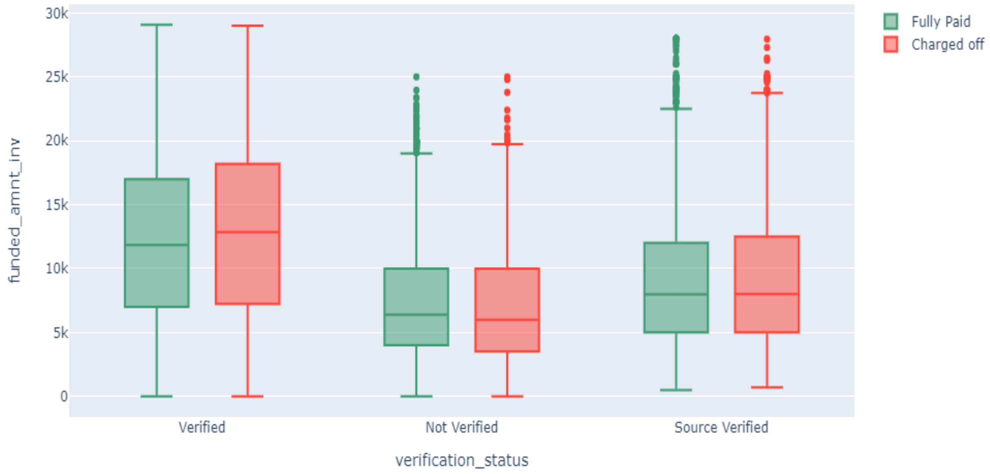


Comparing 25th, 75th percentile between charged off vs fully paid, every loan purpose has slight difference, but for educational this difference is remarkable

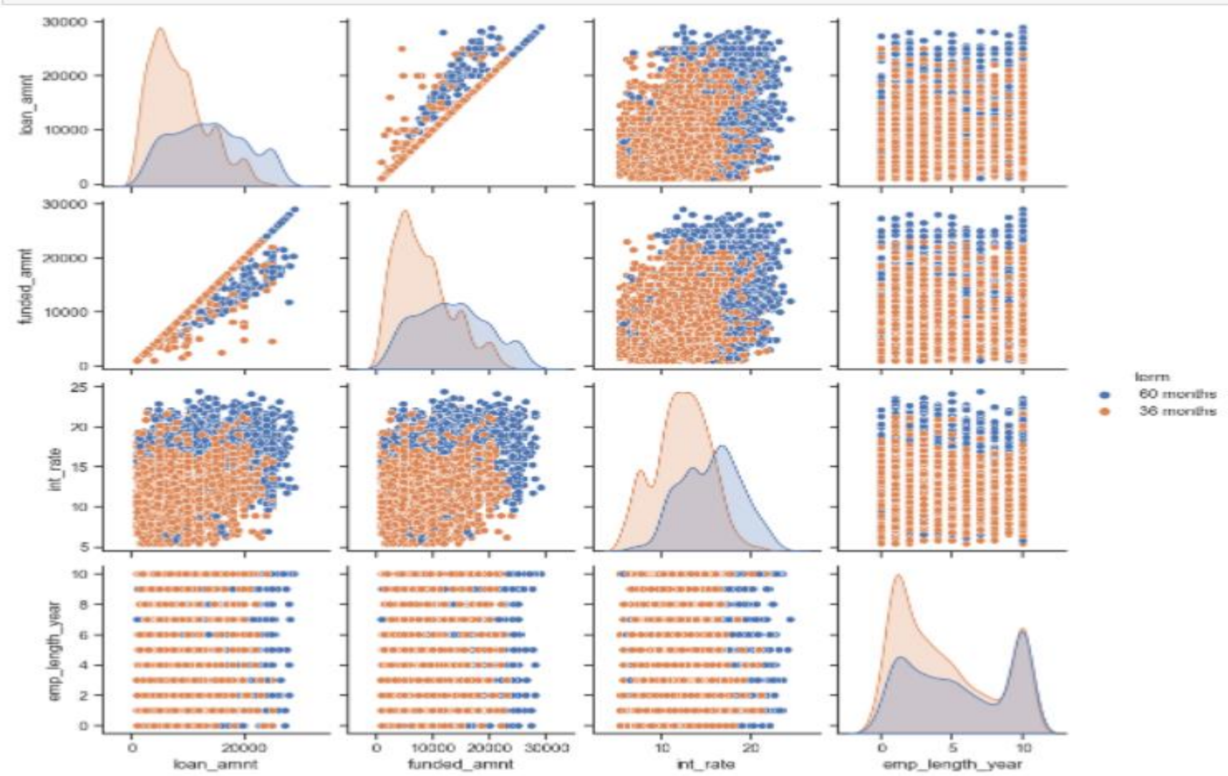
Observation : For educational lower dti also leads to chargeoff, as people taking educational loan may not always be earning

Multi Variate Analysis

Funded Amount Inv is higher if source is verified for all loan status (charged off and fully paid)



Multi-Variate Analysis (Pair Plot)



Conclusions from multivariate

1.Loan amount and funded amount is positively correlated

2.lower interest rate <=18 & funded amount <=20000 were using term 36 months.

3. Higher interest rate :>25 and funded amount upto 30000 were using loan tearm 60 months

4. Irrespective of emp_length , when funded amount > 20000 loan term used is 60 months

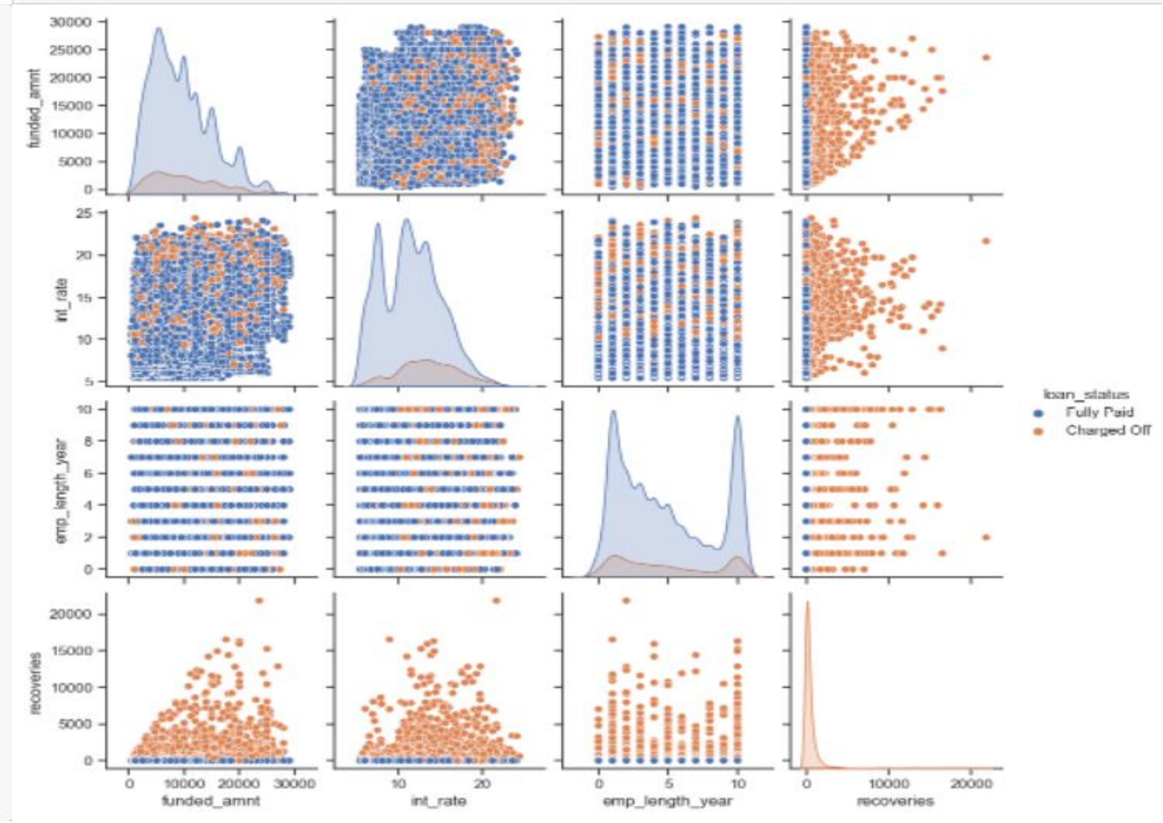
5. Irrespective of emp_length , when funded amount < 20000 term used is 36 months

6. if loan amount > 20000, more 60 months term is observed , loan amount < 20000 36 months term is used

7. if funded amount > 20000, more 60 months is observed , funded amount < 20000 36 months is applied

8. Across all emp length for lower int rate <=18 , 36 months is used. for int rate above 18 , 60 months is observed

Multi-Variate Analysis (Pair Plot)



Conclusions from multivariate

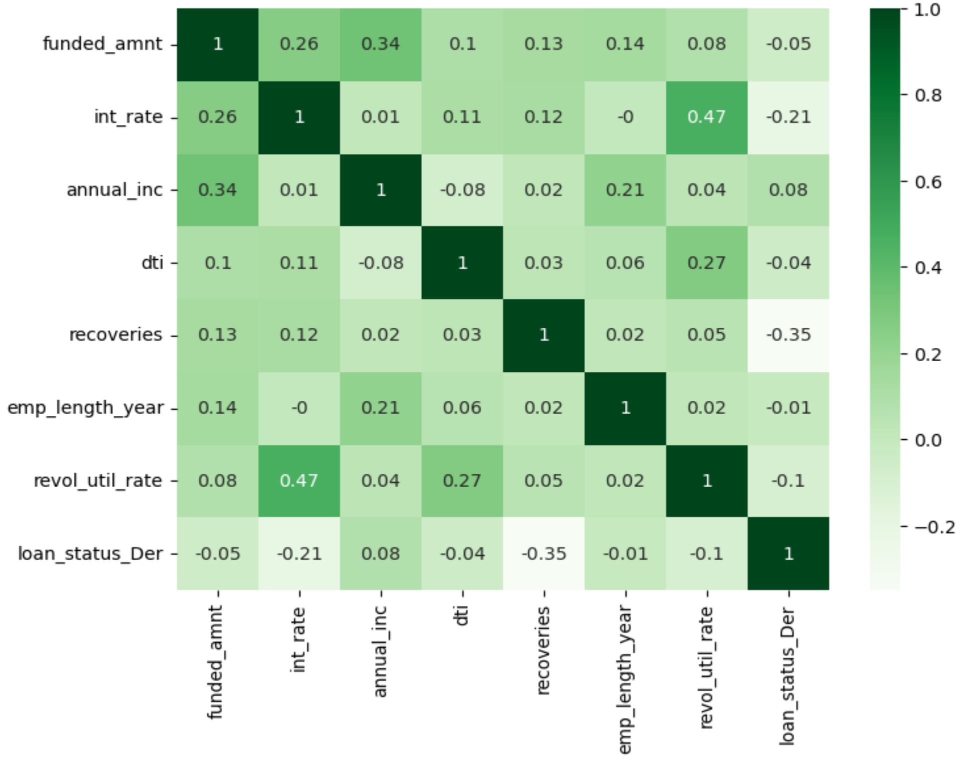
1. Across all emp length with lower int rate till 10 , more is fully paid, Across all emp length once when the int rate > 10 both fully paid and charged off is there

2. Between emp_length vs funded amount , no clean seperation is observed

3. Between int_rate vs funded amount no clean seperation observed between fully paid vs charged off

4. Between int_rate vs recovery , recoveries were zero for fully paid irrespective of int rate . recorveries were higher for charged off acorss all emp length (same applied for all recovery. Recoveries is closely associated to charged off . Not for fully paid)

Multi-Variate Analysis



Conclusions from heatmap

1. loan status towards fully paid is influenced by features like annual income , more higher the annual income more is fully paid

2. loan status towards charged off is influenced by features like

a, funded amount (Higher funded amount lower is fully paid(higher is charged off)

b. int_rate (Higher int rate lower is fully paid(higher is charged off),

c. dti (Higher dti lower is fully paid(higher is charged off),

d. pub rec bankruptcies, ((Higher pub rec bankruptcies lower is fully paid(higher is charged off)

f. revol_util_rate, emp title (Higher revol util rate lower is fully paid(higher is charged off)

Conclusion

Conclusion towards features driving loan defaulters:

1. **Loan Term**: chance of charged off is double when term is 60 months

2. **Purpose** is small business, higher chance of charged off

3. As **grade** increases (A to B to C and so on), higher chance of charged off . Within a grade as **Subgrade** increases (A1 to A5), higher chance of charged off. Chance of charged off is very high if subgrade is F5

4. **Interest rate** - As interest rate increases, % charged off increases

5. Public record **Bankruptcy** is a driver and as this increases, chance of charged off increases

6. **Annual Income**: As Income increases, charged off% decreases, specifically higher chance of charged off if Income less than 60 k

7. **Loan amount** : Higher loan amount increases chance of charge off

8. Emp title : when **emp title is not present**, they have higher chance of charged of

Analysis on multiple features

9. If loan applicant falls in **Grade G and lesser annual Income** chance of charged off is high

10. When **purpose is home improvement/major purpose/moving/small business and income is lower** then higher chance of charged of

Appendix

-

Data Dictionary

Data dictionary: A description and variable definition for all the columns is provided for analysis
https://github.com/shrutich91/LENDING-CLUB-CASE-STUDY/blob/main/Data_Dictionary.xlsx

Data Cleaning

Columns with all null values -54 columns

['mths_since_last_major_derog', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit']

Columns with all same values- 9 columns

['pymnt_plan', 'initial_list_status', 'collections_12_mths_ex_med', 'policy_code', 'application_type', 'acc_now_delinq', 'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens']

Columns with all distinct values -3 columns

['id', 'member_id', 'url']

Columns with > 60% null values -3 columns

['mths_since_last_delinq', 'mths_since_last_record', 'next_pymnt_d']

Cleaning Data – based on data understanding

Customer behavior variables, After studying the data dictionary, further dropping customer behavior variables as it will not provide insights on defaulters

```
["delinq_2yrs","earliest_cr_line","inq_last_6mths","open_acc","pub_rec","revol_bal"
,"total_acc","out_prncp","out_prncp_inv","total_pymnt","total_pymnt_inv","total_rec
_prncp","total_rec_int","total_rec_late_fee","collection_recovery_fee","last_pymnt_
d","last_pymnt_amnt","last_credit_pull_d"]
```

After analyzing description and zip code, since the number of distinct values are very high, it is better to drop these

```
[“desc",“zip_code“]
```